SmartPrevent

Collaborative Project

FP7 - 606952

D4.11 Low-level definition: Methods, techniques and features

Lead Author: Treelogic
With contributions from: VSL
Reviewers: EVS and QMUL

Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	M8
Actual delivery date:	M9
Version:	1.1
Total number of pages:	25
Keywords:	scene understanding, spatio-temporal features, human detection, action recognition

Abstract

In this task we evaluate and select the right set of Computer Vision methods, algorithms and features required for obtaining a rich scene description from raw video frames. As a starting point for the higher-level semantic description of the scene, low-level features are crucial for describing the basic information of the video sequence. By means of the outputs provided by this low-level description layer, each scene, person, activity (and other relevant information) is determined after being captured by the system's visual sensors. Due to the high rates of data to be processed and the performance, robustness and completeness levels expected from the upper layers of the processing pipeline, the set of techniques to be used must be carefully selected. In addition, this task takes into account the requirements and expectations from the layer in charge of developing the action and detection models representing salient actions that can inform the detection of antisocial and/or illegal activities, fed by these low-level features.

[End of abstract]

Executive summary

This document describes the results of the studies performed for identifying, evaluating and selecting the set of low-level features to be computed in the image sequences of video from the visual sensors of the SmartPrevent project.

In SmartPrevent, Treelogic is the partner responsible of the low-level description, and VSL is the partner in charge of performing the high-level semantic description of the scene. Thus, VSL is the consumer of the output from the low-level layer produced by Treelogic. To identify the most suitable methods according to our goals, VSL has provided a comprehensive list of desirable characteristics for this module. For this, they have studied the end-user requirements, previously compiled, and established a bridge between end-user and technical description.

As core part of the report, several state of the art and classical approaches have been studied, based on the results from an extensive literature search, where many authors proposed both traditional and innovative methods for computing those elements of interest. Such methods can be used to describe a visual scene in an automatic, robust and reliable way. Focused on the SmartPrevent project's requirements, we centred the literature review in the following areas:

- Background subtraction - Foreground detection
- Detection and recognition of elements of interest in the scene
- Low-level features of the scene, including:
    - Histograms of Optical Flow (HOF)
    - Histograms of 3D Gradient Orientations (HOG3D)
    - Motion Boundary Histograms (MBH)

Additionally, a summary of relevant validation methods for evaluating the accuracy of the results from this layer is introduced. A more in-depth analysis with appropriate methods will be developed in the upcoming tasks (T6.2 – Evaluation metrics and methods) of the project.

Document Information

IST Project Number	FP7 - 606952	**Acronym**	SmartPrevent
Full Title	Smart Video-Surveillance System to Detect and Prevent Local Crimes in Urban Areas		
Project URL	http://www.SmartPrevent.eu/		
Document URL			
EU Project Officer	Francesco Lorubbio		

Deliverable	**Number**	D4.11	**Title**	Low-level definition: Methods, techniques and features
Work Package	**Number**	WP4	**Title**	Low-Level Scene Understanding

Date of Delivery	**Contractual**	M08	**Actual**	M09
Status	version 1.1		final □	
Nature	prototype □ report X demonstrator □ other □			
Dissemination level	public X restricted □			

Authors (Partner)	Treelogic and VSL			
Responsible Author	**Name**	Víctor Fernández-Carbajales Cañete	**E-mail**	victor.fernandez@treelogic.com
	Partner	Treelogic	**Phone**	+34 910 059 088 (Ext: 5011)

Abstract (for dissemination)	In this task we evaluate and select the right set of Computer Vision methods, algorithms and features required for obtaining a rich scene description from raw video frames. As a starting point for the higher-level semantic description of the scene, low-level features are crucial for describing the basic information of the video sequence. By means of the outputs provided by this low-level description layer, each scene, person, activity (and other relevant information) is determined after being captured by the system's visual sensors. Due to the high rates of data to be processed and the performance, robustness and completeness levels expected from the upper layers of the processing pipeline, the set of techniques to be used must be carefully selected. In addition, this task takes into account the requirements and expectations from the layer in charge of developing the action and detection models representing salient actions that can inform the detection of antisocial and/or illegal activities, fed by these low-level features.
Keywords	scene understanding, spatio-temporal features, human detection, action recognition

Version Log			
Issue Date	**Rev. No.**	**Author**	**Change**
06/10/14	0.1	David Cabañeros Blanco	First draft, including document outline, introduction and abstract.
14/10/14	0.2	David Cabañeros Blanco	Results from studies on low-level scene description methods
14/10/14	0.3	Tao Xiang, Bilal Souti	High-level scene description features
15/10/14	0.4	David Cabañeros Blanco	Merged contributions from both partners. Conclusions
16/10/14	0.5	Víctor Fernández-Carbajales Cañete, Sergio García Álvarez	Final review and improvements

19/10/14	0.6	Zeev Smilansky	Review and edits
21/10/14	0.7	David Cabañeros Blanco, Víctor Fernández-Carbajales	Improvements based on EVS review
27/10/14	0.8	Shaogang Gong	Review and edits
27/10/14	1.0	David Cabañeros Blanco	Improvements based on QMUL review. New version
19/12/14	1.1	David Cabañeros Blanco	Edit and improve the Section 5: Conclusions

Table of Contents

List of figures

List of tables

Abbreviations

BS: Background Subtraction

FD: Foreground Detection

GMM: Gaussian Mixture Model

KDE: Kernel Density Estimation

MOG: Mixture of Gaussians

PCA: Principal Component Analysis

SVM: Support Vector Machine

SVR: Support Vector Regression

ViBe: Visual Background Extractor

GPU: Graphics Processing Unit

1 Introduction

This report describes the specification of the low-level visual features to be used as descriptors to support the semantic description of a scene at the lowest level, so it will represent a key factor for the rest of the SmartPrevent system as a whole.

Although low-level features do not describe a scene nor an activity, they shape the semantic description of a video sequence by providing the required information to the subsequent layer in the video processing pipeline. The results of the work performed in the characterization of the elements composing the scene, including how each one interacts with the rest, will feed the modules developed in WP5. This WP is in charge of understanding criminal activities and suspicious behaviours for helping to predict them, being the latter one of the core objectives of SmartPrevent.

The following list states the key objectives for this task:

1. To study and select a closed set of methods that will be useful for the final system
2. To gather low-level details from the visual scene perception
3. To define the basic elements composing the scenarios of the project, such as people, agglomerations, vehicles, etc.
4. To feed the high-level scene description module with the outputs generated from the low-level feature description module.

In order to establish the group of methods to be applied for obtaining the required features of the video sequences, we have studied, evaluated and selected various state-of-the-art techniques, including background subtraction, tracking, detection and recognition, among others.

2 High-level scene description

As described in the D2.11 Initial Specification and Design of System v1.00 document [1], the SmartPrevent consortium will focus on the following scenarios with different characteristics.

Table 1. Key characteristics for Graffiti Painting scenario

<table>
<tr><th colspan="2">Graffiti Painting</th></tr>
<tr><th>Objectives</th><th>Event Sequence</th></tr>
<tr><td>• Prevent destructive actions against historical heritage (churches, buildings, etc.)
• Prevent citizens to develop negative attitudes against officials</td><td rowspan="2">• One or more people present near a wall
• Carrying backpack
• Carrying a spray can
• People next to a wall
• Wall with new patterns appearing on it
• People filming wall colouring
• People run away</td></tr>
<tr><th>Modules</th></tr>
<tr><td>• Background subtraction
• Person detector
• Actions detector</td><td></td></tr>
<tr><th>Modules Pre-requisites</th><th>Contextual Information</th></tr>
<tr><td>• Real-time
• GPU-based implementation</td><td>• When: nightfall/early morning
• How many: small groups
• How often: quarterly</td></tr>
</table>

Table 2. Key characteristics for Antisocial Behaviour scenario

<table>
<tr><th colspan="2">Antisocial Behaviour</th></tr>
<tr><th>Objectives</th><th>Event Sequence</th></tr>
<tr><td>• Prevent common antisocial behaviour</td><td rowspan="2">• Individual or a group punching or hitting other people
• People throwing different objects
• People gathering and groups creating
• Individual running/jumping</td></tr>
<tr><th>Modules</th></tr>
<tr><td>• Custom objects detector
• Person detector
• Actions detector</td><td></td></tr>
<tr><th>Modules Pre-requisites</th><th>Contextual Information</th></tr>
<tr><td>• Real-time
• GPU-based implementation</td><td>• When: working days evenings/weekends
• How many: crowded
• How often: monthly</td></tr>
</table>

Table 3. Key characteristics for Illegal Parking scenario

<table>
<tr><th colspan="2">Illegal Parking</th></tr>
<tr><th>Objectives</th><th>Event Sequence</th></tr>
<tr><td>• Prevent motor vehicles to park in a non-motor vehicles area
• Increase the safety of the pedestrians in a city centre and very crowded area
• Increase the pedestrian space in a city centre</td><td rowspan="3">• Vehicles driving to a loading/unloading zone and stop
• Vehicles staying more than authorised in a loading/unloading zone</td></tr>
<tr><th>Modules</th></tr>
<tr><td>• Car detector</td></tr>
<tr><th>Modules Pre-requisites</th><th>Contextual Information</th></tr>
<tr><td>• Real-time
• GPU-based implementation</td><td>• When: shops/banks opening hours
• How many: pedestrians crowds
• How often: weekly</td></tr>
</table>

As shown above for these scenarios there are several necessary modules described below.

2.1 Background subtraction module

In order to detect all the foreground objects we need to obtain the scene's static background. This module is needed for two purposes: (1) for the graffiti painting detection, it is needed to detect a permanent change on the wall (the graffiti), which implies that the model needs to have long-term memory so that the changes are not integrated into the background. (2) It will also be used for detecting foreground regions for subsequent tasks including object, person and vehicle recognition and action detection.

2.2 Object detection and tracking

This module will be particularly used for the antisocial behaviour prevention. Indeed apart from detecting person, in order to detect and prevent any throwing or misuse of any objects (bottle, bin, etc...) we need to recognise and detect those objects. Also, to prevent people from parking in pedestrians or illegal areas we basically need to have a car detector in the same way as we have the person detector. In order to do that we need to use some basic low level features (cf. section 3.3) like HOF, HOG, SIFT which will allow us to have the highest possible accuracy in this detection phase. In addition, in order to measure the speed of object movement and extract representation of behaviour based on the trajectory of movement, tracking of each individual objects in the scene is also required.

2.3 Action Detection

The action detection task in complex scenes is quite difficult. In complex scenes, we have complex backgrounds with complete or partial occlusions by crowds; it is very difficult to locate the human body

precisely. In addition, ambiguities may also exist in temporal domain. Since human motion is continuous and since speed varies greatly even within the same action category, it is not easy to decide the start or end point of these actions of interest, even the duration of each action in real world scenarios.

That is why we need a person detector and an action detector both in real time to both detect human bodies in a cluttered environment but also to be able to pinpoint an action like painting on a wall or running away at any given time.

Moreover, as we do not know exactly where and when the target action happens, we need to estimate "a window" covering more than one potential region and time slice. So as the actions are continuous we will not look at one frame at a given time but look at many frames in an interval.

State of the art action detection methods are based on dense trajectory based features and a space-time sliding window search strategy. For real time action detection, extracting features at a very compact region of interest is critical. Therefore optical flow computation is necessary to not only identify the moving regions to focus the computation, but also to identify abrupt change of moving patterns by examining the gradients of flow vectors. In addition, the flow field will be used for features extraction (e.g. the HOF feature).

For real-time application a GPU-accelerated person detector and action feature extraction modules may be needed.

3 Low-level processing pipeline

This section introduces the set of methods to be used for the low-level video sequence understanding layer, including the relationship between them and how they interact to provide a suitable output for the high-level scene description layer.

3.1 Background subtraction

Background subtraction (BS) is a widely used technique in computer vision, specifically in the image processing branch. It is mainly focused on automating the inspection of continuous action scenes, especially within the domain of video surveillance and scene monitoring. The main objective of background subtraction methods is to, given a video sequence, extract those image's regions where the elements of interest are located, e.g. pedestrians, vehicles, abandoned objects, etc.

3.1.1 Working principle

Superficially, background subtraction consists of comparing each input frame from a video sequence against a model that describes the background of the scene, typically composed of static (or almost static, like trees, clouds or whatever type of periodic motion) elements. Once the background is established, we will be able to perform a foreground detection (FD) process for determining which areas of the scene belong to the foreground (referring to those elements not belonging to the scene itself). As a result, a binary mask divides the scene in two classes: foreground (dynamic) and background (static) areas.

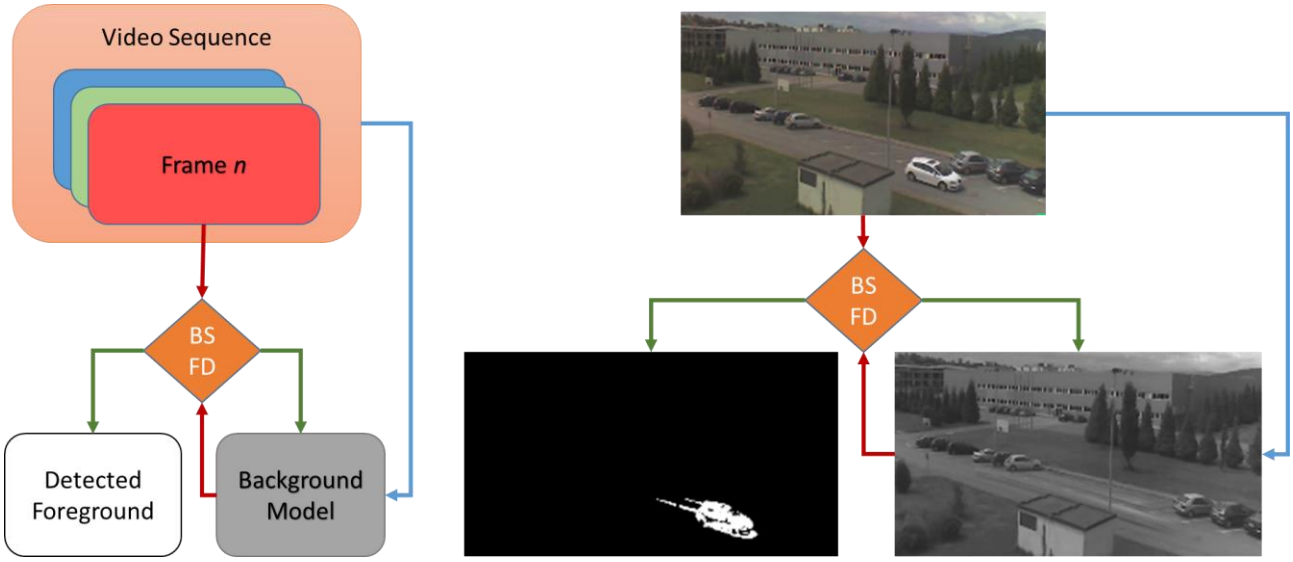

Figure 1. Background Subtraction and Foreground Detection pipeline overview

In SmartPrevent project's context, background subtraction techniques are applied for detecting the scene foreground in order to stand out relevant elements for the low-level semantic description of the sequence. For

this reason, it is mandatory to address relevant problems implicit in outdoor scene analysis, like dynamic backgrounds, shadows or image noise due to stream compression and low resolution sensors.

3.1.2 Considerations and shortcomings

Background subtraction methods are not trivial. This approach has to cope with the inherent particularities of video surveillance. Several authors [2][3][4] evaluated different state of the art approaches, establishing a closed set of common challenges on BS arising from video surveillance applications. The following list enumerates the most common issues to deal with in the field of BS on outdoor environments:

- **Bootstrapping**. Stands for the initialization period where the background of the scene is not available as a static image, so it is difficult to get a clear background representation.
- **Camouflage**. Some elements in the scene may be difficult to distinguish whether they belong to the background or to the background. This situation is usual on the most distant zones of the scene.
- **Dynamic background**. Due to the nature of the scene in outdoor environments, some areas of the image may show movement. Although this motion can be considerate as foreground, it is possible that some of the moving elements should be treated as background, for example, in case of periodical or irregular movements taking place in fixed zones. It is also required to deal with particularities from specific scenarios, such as people drawing graffiti on a wall, where the background of the scene will change permanently.
- **Illumination changes**. Being one of the most relevant issues that complicate the background – foreground detection, we have identified two different variations on the scene lighting.
    - Progressive variation. In outdoor environments the light intensity on the scene varies during day, not only due to the sunlight, but also due to weather conditions (e.g. cloudy days).
    - Sudden variation. Although this change is more related with indoor scenes, lighting may vary due to artificial illumination, such as public lighting, vehicles' headlamps, social events, etc.
- **Image noise**. The image resolution of visual sensors is an important feature to be taken into account. Moreover, video streams are compressed for its transmission, so image artifacts may appear on the picture.
- **Shadows**. Produced by foreground elements itself, shadows may complicate the differentiation of the proper area of the element classified as scene foreground.

This list reflects the main challenges to cope within the background subtraction of the scenes, so techniques to be applied on SmartPrevent low-level scene description module should be selected keeping these potential drawbacks in mind.

Additionally, we must attend to computation time and use of resources for each approach, given that upper layers of the scene description, classification and detection modules will rely on low-level feature extraction performance.

3.1.3 Study of surveyed approaches

In order to evaluate state of the art and conventional techniques for Background Subtraction, the following set of methods, extracted from the literature [5], has been evaluated, attending to its suitability to the previously introduced considerations, and focusing on the speed and robustness required by the online action detection method proposed in the high-level scene description layer.

Due to the numerous approaches found during the study of several surveys [6], we focused on those methods considered more suitable for our needs within the SmartPrevent system, according to the requirements specification.

3.1.3.1 Traditional methods

- **Basic models**. The scene background is computed using average, median or histogram analysis over time. Image pixels are then classified as foreground or as background based on a defined threshold. This is the simplest technique to be applied on BS applications.
- **Statistical models**. Including Gaussian, support vector and subspace learning models.
    - *Gaussian models*. In this way, BS is performed by assuming that the values of each pixel's intensity can be modelled by a Gaussian. Single Gaussian, however, cannot handle dynamic backgrounds, so Mixture of Gaussians (MOG) [7], also referred as Gaussian Mixture Model (GMM), has been adopted as robust and adaptive technique for scenes where dynamic backgrounds are present.
    - *Support vector models*. Based on complex statistical models, like support vector machines (SVM) [8] and support vector regression (SVR) [9]. The former is used to classify all pixels in the image by computing their output probabilities. In the latter models each potential background pixel as an intensity-based function.
    - *Subspace learning models*. The background model is built by applying Principal Component Analysis (PCA) over a set of images. Then, scene foreground is extracted by subtracting this background model from the input image.
- **Cluster models**. Consisting on clustering-based algorithms, these methods assume that each pixel in an image can be represented by clusters. Approaches in this category include K-means [10], Codebooks [11] and basic sequential clustering [12].
- **Estimation models**. Scene background is estimated using a filter that, according to the deviation of each pixel in the image from its predicted value, classifies it into background or foreground classes. Examples of these filters are Wiener [13], Kalman [14] or Chebychev ones.
- **Neural Network models**. These approaches are based on representing the background of the scene by means of the weights of a previously trained neural network, so it can learn how to classify each pixel into background or foreground classes. Several alternatives are included in this category according to the neural network nature.

3.1.3.2 State of the art methods

- **Advanced statistical models**. Including mixture, hybrid and nonparametric models.
    - *Mixture models*. Authors used different distributions apart from the traditional Gaussian approaches, such as Student-t Mixture Model [15] or the Dirichlet Mixture Model [16]. Recently, some approaches have been proposed for reducing the computational cost of these models, looking for a real-time implementation.
    - *Nonparametric models*. These methods are based on following a nonparametric background modelling paradigm. The renowned sample-based algorithm ViBe (Visual Background Extractor) [17] is based on this technique, building the background model by aggregating observed values for each pixel location.
    - *Hybrid models*. Focused on approximate the background colour distribution by means of a combination of a nonparametric regional model (KDE) and a parametric pixel-wise model (GMM).
- **Fuzzy models**. Their purpose is to deal with imprecisions and uncertainties that could appear during the whole process of background subtraction. Several methods [18] for fuzzy background modelling, foreground detection and background maintenance have been proposed.
    - *Fuzzy background modelling*. Consists on modelling multi-modal backgrounds by means of a Gaussian Mixture Model algorithm, complemented with Fuzzy algorithms that provide robustness to the results by modelling uncertainties on dynamic or unstable backgrounds.
    - *Fuzzy foreground detection*. It uses a saturating linear function for contributing to the decision of classifying each pixel into background or foreground.
    - *Fuzzy background maintenance*. In order to update the background model along time, fuzzy foreground detection algorithms are used to decide on the membership (to the foreground or to the background) of each pixel.

3.1.4 Proposed approach

As starting point for the evaluation of suitable BS methods, the early stages of the algorithm development will be implemented using Gaussian Mixture Models (GMM) or Mixture of Gaussians (MOG) based techniques.

Although original GMM/MOG-based background modelling was introduced some years ago, numerous improvements were made along time. Several authors included this type of approach for disparate fields of application, but MOG [19] appears to be the best-performing approach when applied to video surveillance, including detection and recognition.

According to the results of in-depth evaluations found in the comparison performed by Tommesani [20], we conclude that Gaussian Mixture Models implementations offers the most balanced results attending to their computational cost, as shown in the following graph:

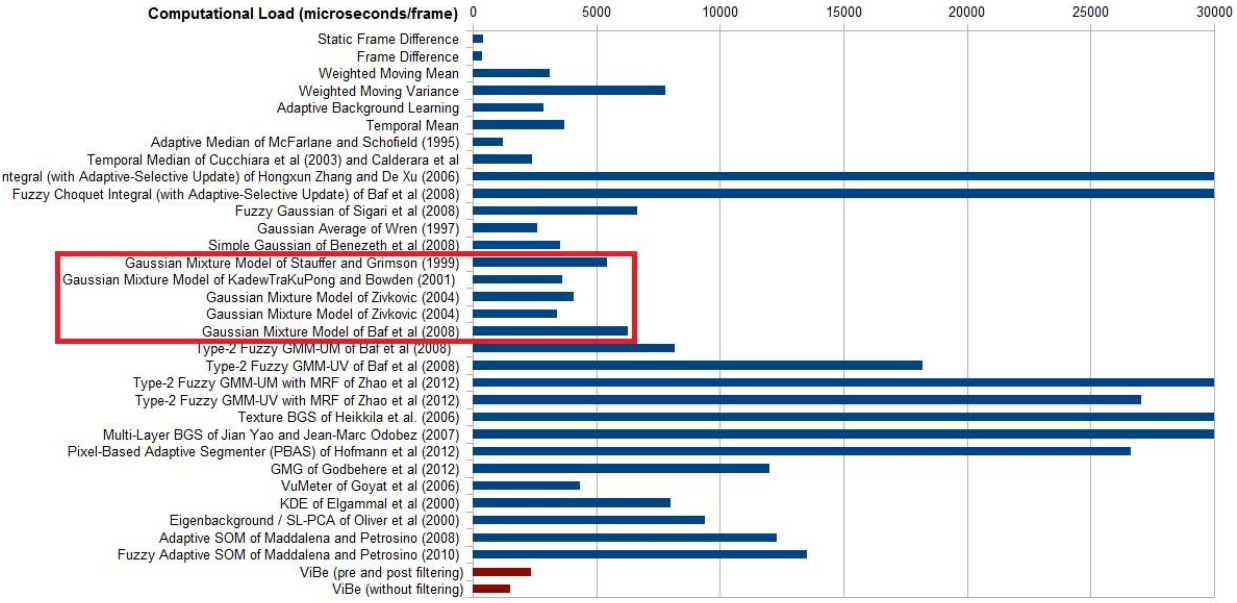

Figure 2. Performance comparison between studied BS techniques

According to its performance, Barnich evaluated their ViBe algorithm against other affine approaches. The results can be compared below:

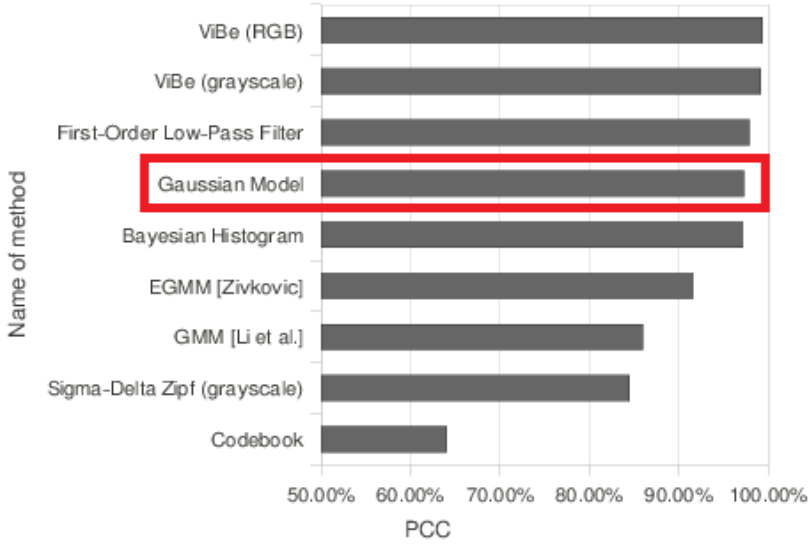

Figure 3. Percentage of Correct Classifications (PCC) for different BS methods

Additionally, the ViBE algorithm has been proposed as an interesting option for this purpose. Due to the restrictions that apply to its usage, its inclusion in the project will be discussed with the project's consortium.

Moreover, taking advantage of the background subtraction algorithm already embedded in the visual sensor to be used in SmartPrevent, we will evaluate its capabilities against the proposed options, in terms of performance, robustness and computational cost. For this purpose, EVS, as supplier of these visual sensors, will provide the required resources and mechanisms for testing these capabilities. This evaluation will also

guide us within the task of proposing and implementing the required improvements for the embedded algorithms included within the sensor.

The main idea is to follow an iterative process where, in collaboration with the developments made on high-level scene description tasks, we will be able to spot and combine those methods, at least, providing the expected results in terms of performance and robustness when applied to the specific focus of the objectives.

3.2 Detection and recognition of elements in the scene

Given that a scene may be composed of several elements, it is necessary to deal with those without any particular interest for semantic description tasks, in order to detect only those zones with relevant visual information for improving the accuracy of the detection. For this reason, a person detector will be implemented with the aim of narrowing down the region where the action to be monitored takes place. Also, for distinguishing between different elements, a contour extraction technique will be used. For this, a binary mask obtained from the detected foreground is used as input. In this way we will be able to clearly establish the exact region of the image corresponding to each segmented element.

The selected method will combine segmentation, detection, recognition and tracking techniques for providing a demarcation of these elements that should be addressed, focusing the usage of computational power required by both low and high level features description algorithms on them.

3.3 Low-level features

In order to compute the required low-level semantic description of the scene, a set of features has been selected, according to their performance and suitability to the expected inputs of the upper layer.

These features are extracted along the sequence's dense trajectories. Dense trajectories are used for describing videos by sampling dense points from each frame of a sequence. Then, these points are tracked based on their displacement information from a dense optical flow field. This procedure is as follows:

1. Densely sample points at several spatial scales
2. Remove the points belonging to homogeneous areas for improving tracking reliability
3. Track points by applying median-filtering in a dense optical flow field
4. Remove static feature trajectories, because they do not contain motion information

Once this trajectory-based representation is achieved, the following descriptors are computed for each trajectory.

3.3.1 HOF – Histogram of Optical Flow

This consists of computing an optical flow sequence and a defining *n* image sub-regions for a temporal reference point, giving a descriptor representing the motion before and after this temporal reference point. This descriptor [21] is known as HOF. Basically, this procedure comprises the following steps:

1. Temporal smoothing of the flow. Based on the temporal median window for a predefined size of sequential images.
2. Discard vectors located outside the region of interest.
3. Split the sequence at the selected temporal reference point into frames before and after this point.
4. Then, for each flow image, the flow field is divided into *n* sub-regions. The 2D histogram of optical flow is calculated for each sub-region at a certain moment.
5. The histogram quantizes flow amplitude and direction.

3.3.2 HOG3D – Histogram of 3D Gradient Orientations

Based on the contributions from Kläser et al [22], HOG3D is proposed as a spatio-temporal descriptor built on the idea behind of how HOG-based descriptors are applied to static images. SmartPrevent's project partners already know about the benefits and performance of combining HOG and HOF approaches. As an innovative method, HOG3D will be implemented as an alternative showing better results. Assuming that HOF performs fast enough, we considered using HOG3D within the scope of this project to find out whether we can afford its more expensive computational cost. Mainly, this descriptor is based on oriented 3D spatio-temporal gradients of each image representing a frame of a video sequence. The process of computing a histogram of 3D gradient orientations is described as follows:

1. Gradient computation. First, gradient vectors are computed for different regions of the image. It is also necessary to take into account several spatial and temporal scaled regions, so these regions will vary both in their positions and boundaries.
2. Orientation quantization. For static (2D) images, an n-bin histogram of gradient orientations is represented as an approximation of a circle with a regular n-sided polygon, where each of these sides correspond to a histogram bin. Applied to 3D images, a polyhedron is used, instead of a polygon, more specifically, the dodecahedron and the icosahedron are considered for 3D gradient quantization.
3. Histogram computation. Given a set of gradient vectors, a histogram of gradient orientations is computed by dividing a block of a region of interest into three sub-blocks. Then, each sub-block's mean gradient is computed and quantized. Finally, the histogram for the region is obtained by summing the quantized mean gradients of all sub-blocks.
4. Descriptor computation. The final descriptor is computed for a local support region around a given sampling point. The width, height and length around the sampling point is parameterized and adjusted.

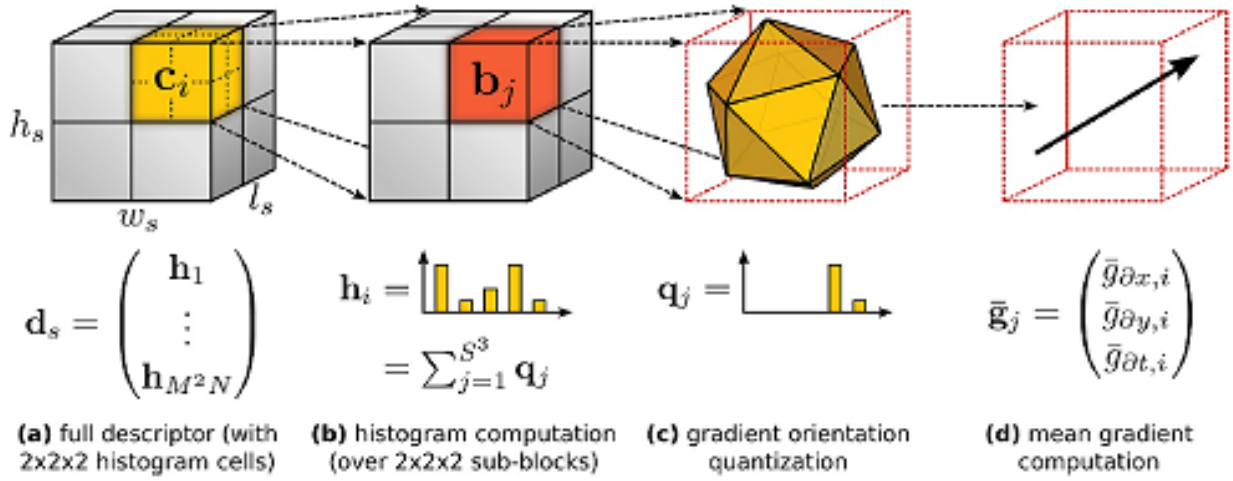

Figure 4 . Overview of the HOG3D descriptor computation process

3.3.3 MBH – Motion Boundary Histograms

MBH-based descriptors rely on the differential of optical flow, outperforming other state-of-the-art descriptors when applied to sequences from real world scenarios [23].

As stated when introducing HOF and HOG3D descriptors, optical flow represents the absolute motion between two different frames, containing motion from different sources appearing in the scene, e.g. cars, people, objects, background, etc. In this way, constant motion information in the scene is supressed, keeping information about changes in the flow field (i.e. motion boundaries).

MBH is employed as a motion descriptor for trajectories. Given the optical flow of an image, MBH descriptors are computed in the following way:

1. Separate optical flow into its horizontal (MBHx) and vertical (MBHy) components
2. Compute spatial derivatives for each plane
3. Quantize the obtained orientation information into histograms
4. Normalize both histogram vectors

After getting this quantization, the orientation magnitude is used for weighting and then, to represent the gradient of optical flow.

Each descriptor is independently computed in each cell of the spatio-temporal grid of the frame. The final descriptor is a concatenation of these descriptors. For HOF and MBH descriptors computation, the dense optical flow already computed for extracting dense trajectories is reused.

4 Validation of the outputs

From the requirements described in Section 2, it is clear that the outputs of the low-level feature modules need to be evaluated based on the following criteria

1. Permanent background change detection accuracy (graffiti on the wall)
2. Detection accuracy for objects
3. Tracking accuracy for objects
4. Movement feature extraction accuracy
5. Real-time performance of all modules

For this evaluation, both subjective and objective validations will be performed. Subjective validation consists of a coarse-grained inspection of the results achieved by both partners involved, checking the outputs of the scene description module by means a human-based visual recognition for assuring the accuracy of the previously described criterions.

For the objective validation, two different options are considered at this stage:

- Validation against external datasets. Several datasets from human action recognition challenges will be studied, attending to their viability for SmartPrevent's specific scenarios.
- Validation against project's datasets. By using the sequences recorded for the project, the accuracy of the outputs from scene description modules will be evaluated.

For both options, the sequences should be annotated in order to be used as ground-truth. Results on accuracy, precision and recall will be extracted from these evaluations.

Task 6.2 from WP6 of the project will provide the final set of evaluation metrics and methods in order to evaluate these outputs.

5 Conclusions

The work performed within this task provided us with the required knowledge about the studied approaches, on which we will rely for establishing a starting point for developing the algorithms. The work carried will be used for the low-level feature extraction of the images composing a video sequence.

We introduced a comprehensive set of key characteristics for each scenario, extracted from the previously performed analysis within the initial specification of the system (WP2). In this way, we are capable to perform an in-depth study of the state of the art methods for focusing the low-level semantic description of the scenes. For example, by taking into account that some events are more likely to take place at specific intervals of the day, we should bear in mind that we have to cope with the variable lighting conditions. Another important issue is the people crowding, common in several scenarios of the project. Section 3.1.2 of this document encloses the main issues affecting the detection of the elements of interest in the scene, being this topic one of the very initial steps of the video processing pipeline.

It results mandatory to carefully choose the most suitable methods from the early phases of the algorithm development, avoiding potential issues to appear once these modules are developed and integrated into the global system's architecture. Studying these methods helped us to *i)* reduce the set of potential methods to be applied and *ii)* select a foreground detection approach that fits the project's requirements defined by the high-level semantic description module. Once the method for extracting the elements of interest in the scene is envisioned, it is necessary to define the procedure for computing the features composing the low-level semantic description of the perceived scene. A hybrid approach, based on a combination of feature descriptors such as HOF, HOG3D and MBH will be used. These descriptors are computed over the dense trajectories generated from the motion of key points of the aforementioned elements of interest appearing in the scene. The decision of using this set of descriptors has been agreed with the partners responsible for developing the layer in charge of, taking these features as input, generating a semantic description of the scene closer to the human cognitive capacities (WP5).

Robustness and performance goals should be also achieved. The scene processing pipeline requires real-time performance for early detection of actions that could lead to criminal activities, so this selection of techniques and their expected refinement and adaptation take a significant role within the project's success.

Through its development, the low-level layer will be validated using the evaluation metrics and methods outlined by the project's test plan, described within WP6. By using both third-party and project's own datasets, several evaluations, focused on different aspects of the outputs generated by this module, will be carried out. The results of this validation process will give us the proof for the capabilities of this module within the overall architecture of the SmartPrevent system.

References

[1] SmartPrevent Deliverable D2.11 “Initial Specification and Design of System”. V1.00

[2] Brutzer, S., Hoferlin, B., & Heidemann, G. (2011, June). Evaluation of background subtraction techniques for video surveillance. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1937-1944). IEEE.

[3] Toyama, K., Krumm, J., Brumitt, B., & Meyers, B. (1999). Wallflower: Principles and practice of background maintenance. In Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on (Vol. 1, pp. 255-261). IEEE.

[4] Bouwmans, T., El Baf, F., & Vachon, B. (2008). Background modeling using mixture of gaussians for foreground detection-a survey. Recent Patents on Computer Science, 1(3), 219-237.

[5] T. Bouwmans, “Traditional and Recent Approaches in Background Modeling for Foreground Detection: An Overview”, Computer Science Review, Volume 11, pages 31-66, May 2014

[6] Maddalena, L., & Petrosino, A. (2008). A self-organizing approach to background subtraction for visual surveillance applications. Image Processing, IEEE Transactions on, 17(7), 1168-1177.

[7] Stauffer, C., & Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on. (Vol. 2). IEEE.

[8] Lin, H. H., Liu, T. L., & Chuang, J. H. (2002, June). A probabilistic SVM approach for background scene initialization. In Image Processing. 2002. Proceedings. 2002 International Conference on (Vol. 3, pp. 893-896). IEEE.

[9] Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. Neural Information Processing-Letters and Reviews, 11(10), 203-224.

[10] Butler, D. E., Bove, V. M., & Sridharan, S. (1900). Real-time adaptive foreground/background segmentation. EURASIP Journal on Advances in Signal Processing, 2005(14), 2292-2304.

[11] Kim, K., Chalidabhongse, T. H., Harwood, D., & Davis, L. (2004, October). Background modeling and subtraction by codebook construction. In Image Processing, 2004. ICIP'04. 2004 International Conference on (Vol. 5, pp. 3061-3064). IEEE.

[12] Xiao, M., Han, C., & Kang, X. (2006, July). A background reconstruction for dynamic scenes. In Information Fusion, 2006 9th International Conference on (pp. 1-7). IEEE.

[13] Chen, J., Benesty, J., Huang, Y., & Doclo, S. (2006). New insights into the noise reduction Wiener filter. Audio, Speech, and Language Processing, IEEE Transactions on, 14(4), 1218-1234.

[14] Welch, G., & Bishop, G. (1995). An introduction to the Kalman filter.

[15] Guo, L., & Du, M. H. (2012, August). Student's t-distribution mixture background model for efficient object detection. In Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on (pp. 410-414). IEEE.

[16] Zivkovic, Z. (2004, August). Improved adaptive Gaussian mixture model for background subtraction. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 2, pp. 28-31). IEEE.

[17] Barnich, O., & Van Droogenbroeck, M. (2011). ViBe: A universal background subtraction algorithm for video sequences. Image Processing, IEEE Transactions on, 20(6), 1709-1724.

[18] Bouwmans, T. (2012). Background subtraction for visual surveillance: A fuzzy approach. Handbook on Soft Computing for Video Surveillance, 103-134.

[19] Wang, H., & Suter, D. (2005, March). A re-evaluation of mixture of Gaussian background modeling [video signal processing applications]. In Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP'05). IEEE International Conference on (Vol. 2, pp. ii-1017). IEEE.

[20] Tomessani, S. Comparing background subtraction algorithms. Online. Fetched 20th October, 2014 - http://tommesani.com/index.php/video/comparing-background-subtraction-algorithms.html

[21] Perš, J., Sulić, V., Kristan, M., Perše, M., Polanec, K., & Kovačič, S. (2010). Histograms of optical flow for efficient representation of body motion. Pattern Recognition Letters, 31(11), 1369-1376.

[22] Klaser, A., & Marszalek, M. (2008). A Spatio-Temporal Descriptor Based on 3D-Gradients.

[23] Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013). Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1), 60-79.